RANKA
DATE / /
PAGE

STATISTICS

⇒ Statistics → Statistics is the science of collecting, organizing and analyzing the data.

Data → facts or pieces of information.

STATISTICS

Descriptive Stats (EDA + FE)	Inferential Stats
① It consists of organizing and summarizing the data.	It consists of collecting sample data and making conclusion about population data using some experiments.
FE → Feature Engineering	

Sampling Techniques

1) Simple Random Sampling – Every member of the population (N) has an equal chance of being selected for your sample(n)
Eg:- Exit Poll, General survey, Movie reviews

2) Stratified Sampling:- Strata (layers) → Clusters → Groups-
Eg:- Based on Gender, Blood Group.

3) Systematic Sampling:- Select every n^{th} individual out of population (N).

4) Convenient Sampling – Only those who are interested in the survey will only participate.

* Variable → A variable is a property that can take any values

i) Quantitative Variable → Measured Numerically {Mathematical operations can be done} Eg- Age, weight, height, rainfall (cm), temp, distance.

ii) Qualitative variable :- Categorical variables {Based on some characteristics they are grouped together}

Eg:- Gender, Types of flowers, Types of Movies.

Quantitative Variable

↓ Discrete Variable

Eg:- Whole numbers only
No. of bank accounts, children in family
Pincode is an example. Also there can be many numbers of data points.

↓ Continuous Variable

Decimal numbers
Eg:- Height, weight, ages, rainfall, speed.

Histograms

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

1) Sort the numbers
2) Bins → No. of groups
3) Binsize.

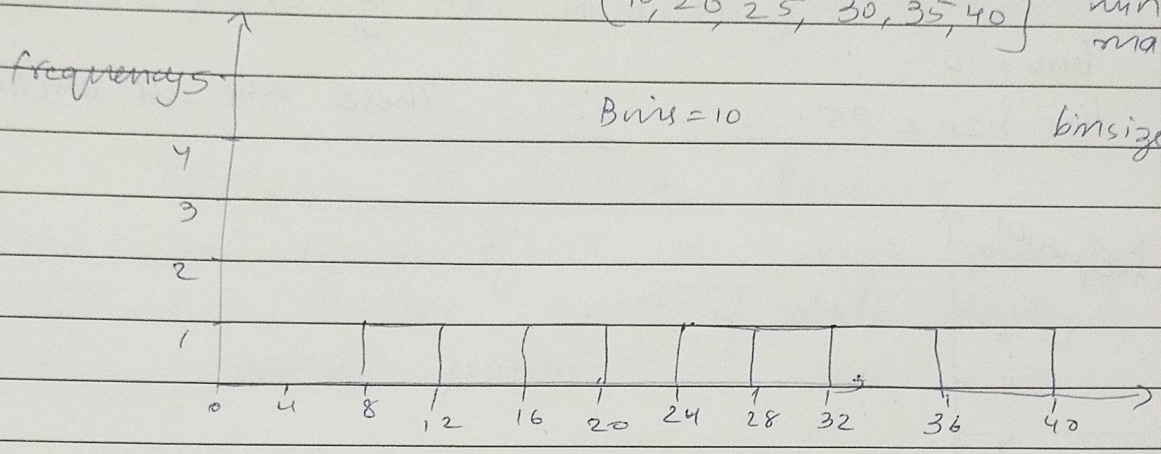
RANKA
DATE / /
PAGE

lets, look at an example.

[10, 20, 25, 30, 35, 40] min = 10, max = 40

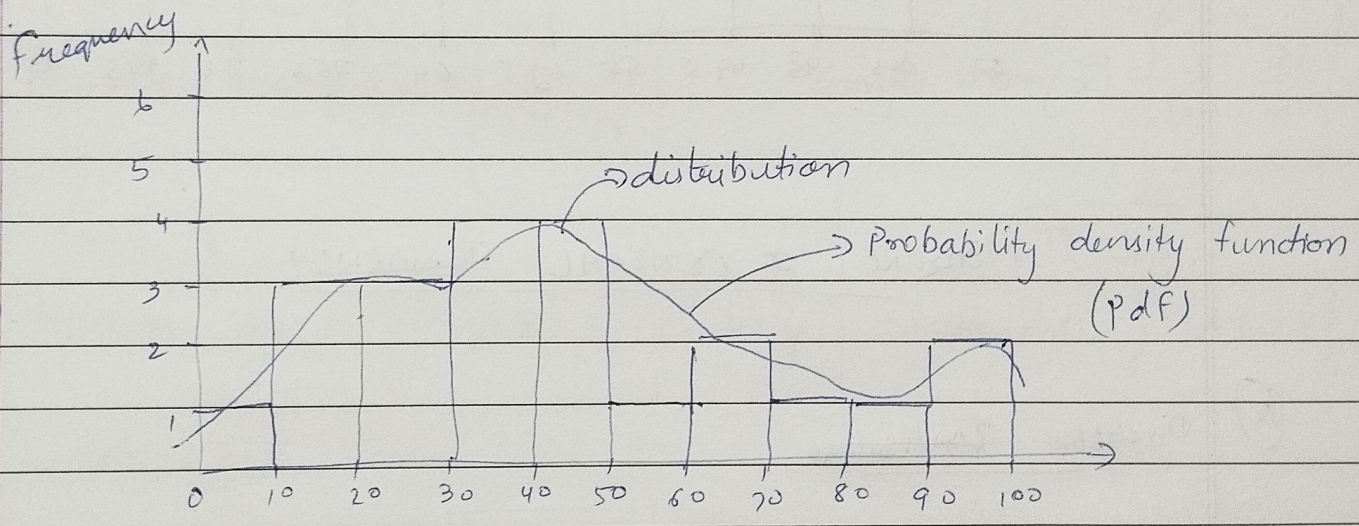
Bins = 10

$binsize = \frac{Max}{Bins} = \frac{40}{10}$



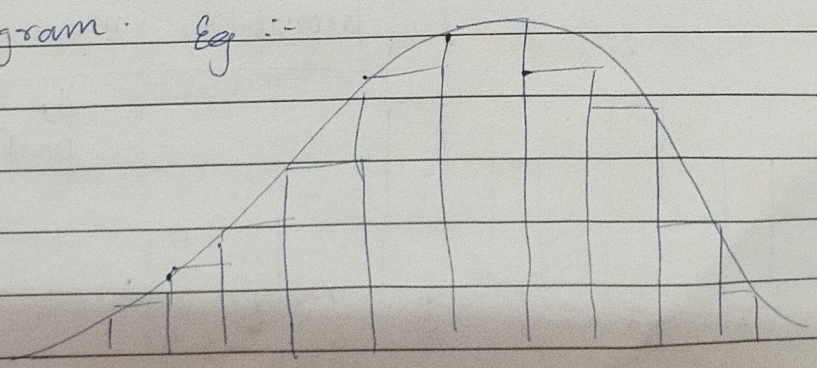
* for Ages. $Binsize = \frac{100}{10} = \frac{10}{}$

10 → Assumed



* We are smoothening the curve, this shows how the data is getting distributed.

* When we smoothen the curve internally there will be histogram. Eg :-

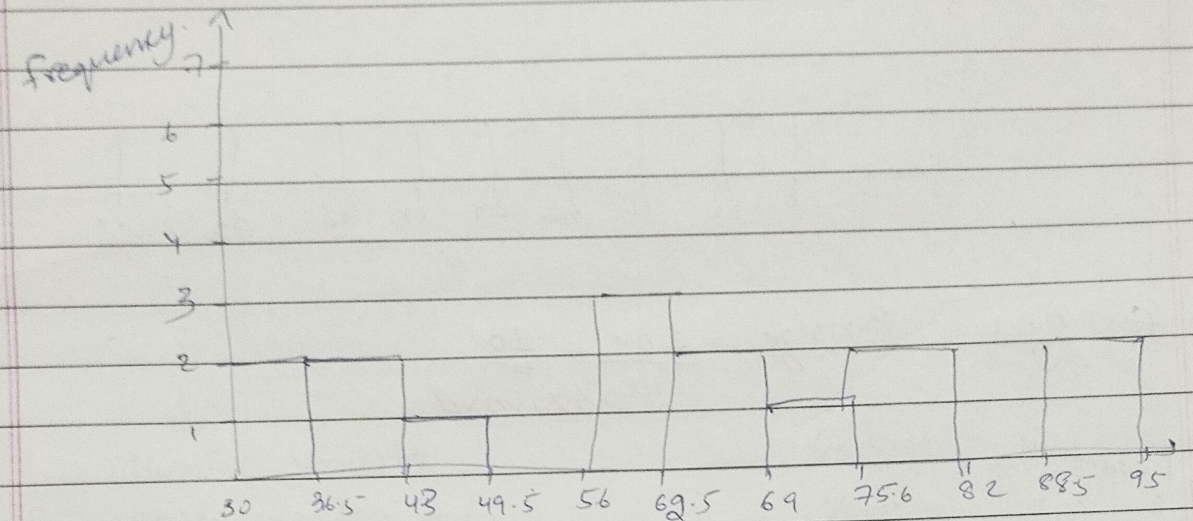


Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

↓ These are all continuous value

bins = 10

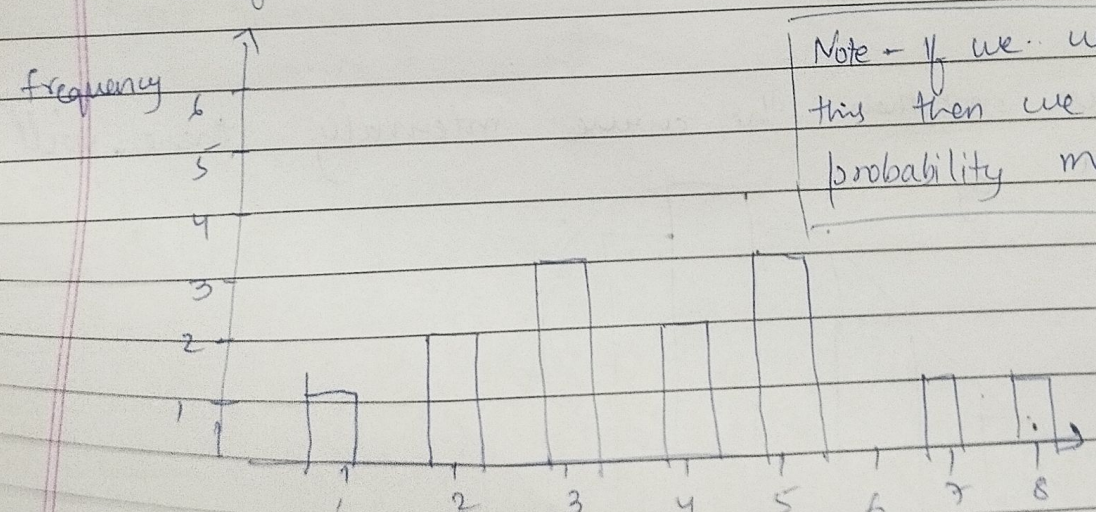
bin size = $\frac{95-30}{10} = 6.5$



MEASURE of CENTRAL TENDENCY

(a) Discrete ~~Continuous~~

No. of bank accounts = {2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5}



Note - If we will smoother this then we have to use probability mass function

↓

pmf

pdf : probability density function } → continuous
pmf : probability mass function } → discrete

MEASURE OF CENTRAL TENDENCY

① Mean
② Median
③ Mode

A measure of CT is a single value that attempts to describe a set of data identifying the central position.

Mean $X = \{1, 2, 3, 4, 5\}$, Average/Mean $= \frac{15}{5} = 3$

Population (N) $\quad N >> n \quad$ Sample (n)

Population mean (μ) $= \frac{\sum_{i=1}^{N} x_i}{N}$ $\qquad$ Sample mean ($\bar{x}$) $= \frac{\sum_{i=1}^{n} x_i}{n}$

$N = 6$ $\qquad$ $n = 4$

Population Age $= \{24, 23, 21, 28, 27\}$ $\qquad$ Sample Ages $= \{24, 21, 27\}$

$\mu = 17.5$ $\qquad$ $\bar{x} = 13.5$

$\mu \geq \bar{x}$

$\bar{x} \geq \mu$

Practical Application (Feature Engineering)

Age	Salary	family size
—	—	—
nan	—	—
—	nan	—

We can use mean in nan.

② Median

$\{1, 2, 3, 4, 5\}$ = $\{1, 2, 3, 4, 5, \boxed{100}\}$ ← Outlier

$\bar{x} = 3$ ⟶ $\bar{x} = 19.16$

Steps to find out median

1. Sort the numbers
2. Find the central number
   - i) if the no. of elements are even we find the average of central elements.
   - ii) if the no. of elements are odd we find the central elements.

Sorted

$\{1, 2, 3, 4, 5, 6, 7, 8, 100, 120\}$

Median = $5.5 = \boxed{\frac{5+6}{2}}$ Mean = 25.6

③ Mode = Most frequent occurring element.

$\{1, 2, 2, 3, 3, 3, 4, 5\}$ ⟶ Mode = 3

$\{1, 2, 2, 2, 3, 3, 3, 4, 5\}$ ⟶ Mode = $\boxed{2, 3}$

* Measure of Dispersion

1. Variance (σ^2) ← Spread of Data
2. Standard deviation (σ)

RANKA
DATE / /
PAGE

Variance

Population variance (σ^2)

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$\{1, 2, 3, 4, 5\}$

$\mu = 3$

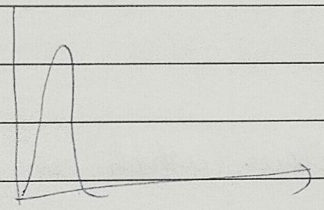
Sample variance (s^2)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

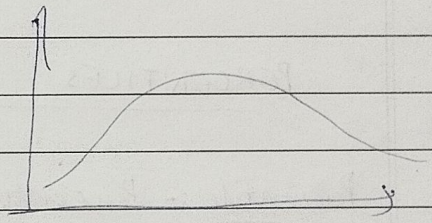
$\{1, 2, 3, 4, 5, 6, 80\}$

$\mu = 14.4$

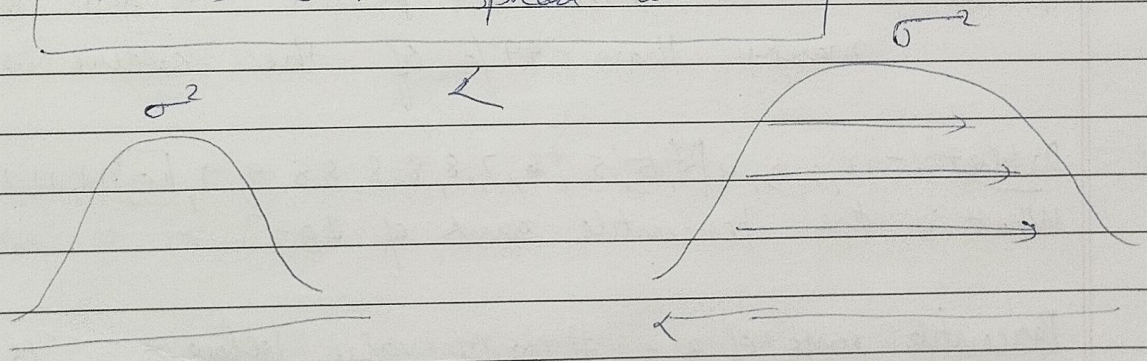
$\sigma^2 = 2$



$\sigma^2 = \uparrow\uparrow$



Note – As $\sigma^2 \uparrow$ spread also $\uparrow\uparrow$



Q. Why sample variance is divided by $n-1$.

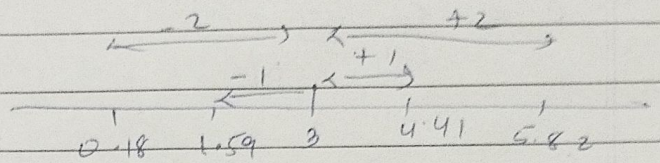
* Standard deviation ($\sqrt{\sigma^2}$)

$\{1, 2, 3, 4, 5\}$ $\}1$

$\mu = 3$

$\sigma^2 = 2$

$\sigma = 1.41$



Standard deviation indicates how many standard deviation the data points are away from the mean.

PERCENTILES AND QUARTILES

Percentile :- Percentile is a value below which a certain percentage of observation lie.

99 percentile ⟹ It means the person has got better marks than 99% of the entire students.

Dataset :- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

What is the percentile rank of 10.

Percentile rank of x = $\frac{\#\text{ No. Of value below } x}{n}$ = $\frac{16}{20} = \frac{4}{5} = 80\%$

* What is the value that exists at 25 percentile.

$0.25 = \frac{n}{20}$

$n = 5^{th}$ → index (5 is the value)

in dataset consider from 0.

$$\text{Value} = \frac{\text{Percentile} \times n}{100}$$

* 5 Number summary

1) Minimum
2) First Quartile (25 percentile) Q_1
3) Median
4) Third Quartile (75 percentile) (Q_3)
5) Maximum

{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

[Lower Fence ⟷ Higher Fence]

75% 25 per

$IQR = Q_3 - Q_1$

⇓

Inter Quartile Range

Lower Fence $= Q_1 - 1.5\,(IQR)$

Higher Fence $= Q_3 + 1.5\,(IQR)$

↓

1.5 Standard Deviation

$Q_1 = \frac{25}{100} \times (n+1) = 5.25$ index

Calculate Avg of 5th & 6th index is 3 because 5.25 index is no there

$Q_3 = \frac{75}{100}(n+1) = 0.75 \times 21 = 15.75$ index

Calculating avg of 15th & 16th index → 7.5

$Q_1 = 3$

$Q_3 = 7.5$

$IQR = 4.5$

Lower fence $= 3 - (1.5) \times 4.5 = -3.65$

Higher fence $= 7.5 + 1.5 \times 4.5 = 14.25$

$[-3.65 \longleftrightarrow 14.25]$

This means in the dataset 27 is an outlier.

{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

(1) Min = 1
(2) $Q_1 = 3$
(3) Median = 5
(4) $Q_3 = 7.5$
(5) Max = 9